Simon, Thorsten; Fabsic, Peter; Mayr, Georg J.; Umlauf, Nikolaus; Zeileis, Achim

Working Paper

Probabilistic forecasting of thunderstorms in the Eastern Alps

Working Papers in Economics and Statistics, No. 2017-25

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Simon, Thorsten; Fabsic, Peter; Mayr, Georg J.; Umlauf, Nikolaus; Zeileis, Achim (2017) : Probabilistic forecasting of thunderstorms in the Eastern Alps, Working Papers in Economics and Statistics, No. 2017-25, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/184974

universität innsbruck

Research Platform
Empirical and Experimental Economics

working paper

eeecon
[triple:e:con]

Probabilistic Forecasting of Thunderstorms in the Eastern Alps

Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis

Working Papers in Economics and Statistics

2017-25

University of Innsbruck
http://uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

Probabilistic Forecasting of Thunderstorms in the Eastern Alps

Thorsten Simon
University of Innsbruck

Peter Fabsic
University of Innsbruck

Georg J. Mayr
University of Innsbruck

Nikolaus Umlauf
University of Innsbruck

Achim Zeileis
University of Innsbruck

Abstract

A probabilistic forecasting method to predict thunderstorms in the European Eastern Alps is developed. A statistical model links lightning occurrence from the ground-based ALDIS detection network to a large set of direct and derived variables from a numerical weather prediction (NWP) system. The NWP system is the high resolution run (HRES) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The statistical model is a generalized additive model (GAM) framework, which is estimated by Markov chain Monte Carlo (MCMC) simulation. Gradient boosting with stability selection serves as a tool for selecting a stable set of potentially nonlinear terms. Three grids from 64×64 km^2 to 16×16 km^2 and 5 forecasts horizons from 5 to 1 day ahead are investigated to predict thunderstorms during afternoons (1200 UTC to 1800 UTC). Frequently selected covariates for the nonlinear terms are variants of convective precipitation, convective potential available energy, relative humidity and temperature in the mid layers of the troposphere, among others. All models, even for a lead time of five days, outperform a forecast based on climatology in an out-of-sample comparison. An example case illustrates that coarse spatial patterns are already successfully forecast five days ahead.

Keywords: lightning detection data, statistical post-processing, generalized additive models, gradient boosting, stability selection, MCMC.

1. Introduction

Predicting thunderstorms in complex terrain is a challenging task since one of the main tools, numerical weather prediction (NWP) systems, cannot fully resolve convective processes nor circulations and exchange processes over complex topography. Thus NWP output is statistically post-processed to enhance its value for thunderstorm forecasts. Logistic regression is often employed for predicting whether thunderstorms will occur (Schmeits *et al.* 2008; Gijben *et al.* 2017).

However, two difficulties are present: Firstly, the response variable (probability of thunderstorms) might nonlinearly depend on individual covariates from the NWP. Secondly, an abundance of potential covariates provided by NWP systems could be included in the statis-

tical model. Thus a statistical framework capable of handling nonlinear relationships between the response and covariates and objectively selecting the important covariables is needed.

Nonlinearities can be captured either by transformations of covariates, e.g., power or log transformation, or by nonlinear regression models, e.g., a *generalized additive model* (GAM, Hastie and Tibshirani 1990; Wood 2017). GAMs can be formulated in Bayesian framework (Brezger and Lang 2006) which allows to estimate GAMs using Markov chain Monte Carlo (MCMC) simulations. This approach is in particular attractive for inference of complex GAMs (Umlauf *et al.* 2017). GAMs have been used for post-processing NWP output to capture complex spatio-temporal characteristics of temperature (Dabernig *et al.* 2017) and precipitation (Stauffer *et al.* 2017).

Selection is classically performed by testing all possible subsets of potential covariates (Miller 2002). This procedure becomes computationally intractable for large numbers of covariates as in our case. Thus non-exhaustive methods such as stepwise selection are more common (Miller 2002). In recent years also regularization methods have become popular for variable selection in the field of post-processing NWP output, e.g., the LASSO (Wahl 2015) and boosting (Messner *et al.* 2017).

Gradient boosting was first established in the field of machine learning (Freund and Schapire 1995), and generalized later by Bühlmann and Hothorn (2007) for regression models such as GAMs. A broad overview of algorithms for this technique can be found in Mayr *et al.* (2012). However, selecting the right-sized subset of covariables remains challenging (Meinshausen and Bühlmann 2010), i.e., to avoid selecting some noise variables (Hofner *et al.* 2015). A solution to this issue is combining gradient boosting as a method of regularization with stability selection (Hofner *et al.* 2015).

The aim of this study is to develop a probabilistic forecasting method for the occurrence thunderstorms in the Eastern Alps and their surroundings. In order to achieve this objective we propose a novel combination of the statistical methods introduced above. A GAM serves as framework to account for potentially nonlinear relationships between response and covariates. Within this framework an objective variable selection scheme, i.e., gradient boosting with stability selection, is performed to select a stable subset of the available covariates. In a final step the GAM comprising the selected terms is estimated using MCMC sampling. This allows to draw inferential conclusions such as credible intervals of effects, predictions or out-of-sample scores.

The region we focus on are the Eastern Alps in Europe and their surroundings (Figure 1). The region is exposed to severe thunderstorms and lightning during summer (Schulz *et al.* 2005; Simon *et al.* 2017). Furthermore the Eastern Alps are characterized by a complex terrain. Elevation within the study domain extends from sea level up to 3798 meter a.m.s.l. The atmospheric processes leading to the strong convective events and the occurrence of thunderstorms in this region cover the gamut from small to large scales. Interactions of orography, solar heating and winds influence the lightning activity (Bertram and Mayr 2004; Houze 2014). On the other hand, large scale circulations, e.g., the North Atlantic Oscillation, might influence the lightning patterns in Europe (Piper and Kunz 2017). Studies investigating the climatological patterns of lightning activity in the region of interest and its surrounding found maxima along the northern and southern rim of the Alps (Schulz *et al.* 2005; Feudale *et al.* 2013; Wapler 2013).

This manuscript is structured as follows: The region of interest, the lightning detection data

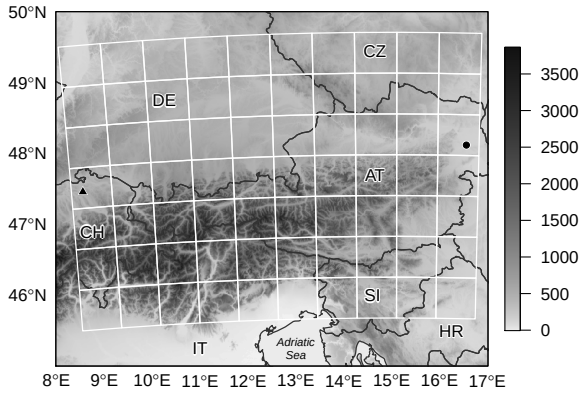

Figure 1: Topography of the region of interest [m a.m.s.l.]. The white solid lines indicate the 64×64 km^2 spatial grid. The circle and the triangle show the location of the Zurich airport (ZRH) and Vienna airport (VIE), respectively.

and the covariates are presented in Section 2. The statistical methods—GAMs, gradient boosting with stability selection and MCMC—are introduced in Section 3. In Section 4 the results of the selection scheme are presented in detail for one model and the predictive performance of the models across different spatial and temporal scales is analyzed. An example forecast is discussed in Section 5. Finally, the study is summarized and concluded in Section 6.

2. Data

In the following the response variable based on lightning detection data, and the covariates from the ECMWF high resolution run are introduced. The study covers the times when most thunderstorms occur (Bertram and Mayr 2004; Schulz *et al.* 2005; Simon *et al.* 2017), i.e., the afternoons (1200–1800 UTC) of the convective season May–August of the years 2010–2015, during which the horizontal mesh of the ECMWF high resolution run remained unchanged at 16 km. The region we focus on are the Eastern Alps in Europe and their surroundings (Fig. 1). The region is horizontally divided into multiples of the ECMWF grid to study the dependence of the forecast performance on spatial resolution. The three meshes are 64×64 km^2, 32×32 km^2 and 16×16 km^2. The grid with the coarsest spatial resolution—64×64 km^2—is highlighted in Figure 1 by white solid lines.

Thunderstorms based on lightning detection data

Thunderstorms are set to have occurred in a grid cell when at least one lightning stroke was registered by the ground-based ALDIS lightning detection network (Schulz *et al.* 2005) between 1200 and 1800 UTC.

The resulting sample sizes of the data sets for the three resolutions 64×64 km^2, 32×32 km^2 and 16×16 km^2 are 51660, 221400 and 885600 with the unconditional probability of lightning activity of 30.3%, 19.7% and 12.5%, respectively.

Covariates from the ECMWF high resolution run

The covariates for predicting the occurrence of lightning activity are derived from the high resolution run (HRES) from the European Centre for Medium-Range Weather Forecasts (ECMWF) initialized at 0000 UTC. The horizontal mesh of 16×16 km^2 remained unchanged during the study period 2010–2015. A list of variables selected for this study is given in Table 1.

Table 1: An overview of the base covariates from the ECMWF-HRES forecast. Covariates derived from this base set are discussed in the data section.

Abbreviation	Description
`sqrt.cp`	Square root of convective precipitation.
`sqrt.cape`	Square root of convective available potential energy.
`t500`, `t700`, `t2m`	Temperature at 500 hPa, 700 hPa, and 2 meters.
`d2m`	Dew point temperature at 2 meters.
`r700`	Relative humidity at 700 hPa.
`w500`, `w700`	Pressure vertical velocity at 500 hPa and 700 hPa.
`ws700`	Wind speed at 700 hPa.
`wdir700`	Wind direction at 700 hPa.
`mls`	Proxy for mid-layer stability: $\texttt{mls} \propto \texttt{t500} - \texttt{t700}$.
`tcc`	Total cloud cover.
`slhf`	Surface latent heat flux.
`ssr`	Surface net solar radiation.
`str`	Surface net thermal radiation.
`e`	Evaporation.

The variables are prepared for the lead times 12h/15h/18h (day 1), 36h/39h/42h (day 2), 60h/63h/66h (day 3), 84h/87h/90h (day 4) and 108h/111h/114h (day 5), which are used to build the base for five different models with respect to each resolution.

Additional covariates are derived from the variables in Table 1. The mean, maximum and minimum of the values at 1200 UTC, 1500 UTC and 1800 UTC of a specific variable are denoted by the name of the variable and the suffix `.mean`, `.max` and `.min`, respectively. Differences between different times are marked by suffices with four digits, e.g., `t700_1812` is the temperature difference at 700 hPa between 1800 and 1200 UTC. The first and the last two digits of the suffix correspond to different times. Finally, anomalies computed by subtracting the mean values are marked by suffices with two digits, e.g., `t700_12` for the temperature anomaly at 700 hPa at 1200 UTC. This procedure leads to a total of 126 potential covariates derived from the NWP model.

3. Methods

In this section the framework of *generalized additive models* (GAMs), which allows for modeling potentially nonlinear smooth functions of the covariates, is described. Furthermore, we give an explanation how variable selection is performed using *gradient boosting* with *stability selection*, and how inference of the finally selected model is made by *Markov chain Monte Carlo* (MCMC) sampling.

Generalized additive models

The statistical framework to model lightning activity falls in the class of *generalized additive models* (GAM). A comprehensive introduction to GAMs is given by Wood (2017).

The dichotomous variable of observing *lightning* or *no lightning* activity in a grid cell follows a Bernoulli distribution with the parameter π which is the probability of observing lightning activity. The logit function links π to an additive predictor η,

$$\text{logit}(\pi) = \eta = \beta_0 + f_{\text{time}}(\texttt{doy}) + f_{\text{space}}(\texttt{lon}, \texttt{lat}) + f_3(\texttt{x}_3) + \cdots + f_p(\texttt{x}_p). \tag{1}$$

The intercept in the additive predictor is β_0, f_j are potentially nonlinear functions modeled here by P-splines (Wood 2017), and the covariates $\texttt{x}_j$ are derived from the ECMWF-HRES (Table 1).

Two further additive terms account for seasonal and spatial variations: f_{time} depends on the *day of the year* (`doy`) and f_{space} on *longitude* (`lon`) and *latitude* (`lat`). Thus the total number of potential terms of the GAM is $p = 128$.

The first three components of Eq. 1—intercept, temporal and spatial effect—are employed to build a baseline model that describes the climatological probability of lightning,

$$\text{logit}(\pi_{\text{baseline}}) = \beta_0 + f_{\text{time}}(\texttt{doy}) + f_{\text{space}}(\texttt{lon}, \texttt{lat}). \tag{2}$$

The response variable follows a Bernoulli distribution with the associated log-likelihood function,

$$\ell(\pi) = y \cdot \log(\pi) + (1 - y) \cdot \log(1 - \pi), \tag{3}$$

for an individual observation $y \in \{0, 1\}$.

To ensure regularization of the functions f_j and to prevent overfitting, in the frequentist approach so called penalty terms are added to the objective log-likelihood function such that the smoothness of each function is controlled by additional smoothness parameters which need to be estimated, e.g., by additionally minimizing the AIC or by restricted maximum likelihood (REML, Wood 2017). In boosting, the smoothness parameters are utilized to initialize each function f_j with the same degrees of freedom to ensure an equal comparison for the selection of base-learners in each boosting iteration (Bühlmann and Hothorn 2007). The Bayesian analogue of the frequentist penalty terms are shrinkage priors that are assigned to the corresponding regression coefficients of each function f_j. These priors are commonly based on multivariate normal priors (Umlauf *et al.* 2017). Hence, the regression coefficients and the smoothness parameters can be estimated simultaneously using MCMC sampling.

In this study we propose a novel combination of methods in order to obtain a final GAM. First, *gradient boosting with stability selection* serves for selecting a stable subset of terms. Second, the selected model is estimated using *MCMC* sampling which allows drawing inferential conclusions about the selected terms.

Gradient boosting with stability selection

The selection of informative nonlinear functions f_j is performed by gradient boosting (Mayr *et al.* 2012) combined with stability selection (Meinshausen and Bühlmann 2010).

Gradient boosting is an iterative gradient descent algorithm, where the term which fits best to the gradient of the log-likelihood is slightly updated in each iteration. The iteration steps are:

1. Initially all terms (or *base-learners*) are set equal to zero, i.e., $f_j(\mathbf{x}_j) = 0$.

2. In each iteration k, the negative gradient of the log-likelihood $-\partial\ell/\partial\eta^k$ is evaluated for every observation, leading to a vector of gradients.

3. For each term $f_j(\mathbf{x}_j)$, low-degree-of-freedom splines are fitted to the gradient vector using penalized least squares estimation.

4. The coefficients of the best fitting term—with respect to the residual sum of squares—are updated by a proportion ν, here $\nu = 0.1$, leading to an updated predictor,

$$\eta^{k+1} = \eta^k + \nu \cdot f_j(\mathbf{x}_j). \tag{4}$$

5. Steps (2–4) are repeated for a predefined number of iterations $k_{\max}$ or until a predefined number of terms q has been selected.

If gradient boosting is applied as stand-alone method the number of iterations $k_{\max}$—and thus the degree of regularization—can be determined by means of information criteria or cross-validation. Here the main purpose of gradient boosting is selecting important terms f_j. It is desirable to avoid the selection of numerous non-informative terms. Stability selection is a convenient resampling method for controlling the number of selected non-informative terms by gradient boosting (Meinshausen and Bühlmann 2010; Hofner *et al.* 2015).

Rather than applying this boosting approach to all n observations, stability selection is based on drawing a subsample of size $n/2$ from the training data, running the boosting algorithm until q base learners are selected. This procedure is repeated many times. Afterwards the relative selection frequencies per base learner are computed. Eventually the base-learners for which the relative selection frequency exceeds a certain threshold are included in the final model (cf. algorithm in Hofner *et al.* 2015).

Markov chain Monte Carlo sampling

The final model is of a complex form as it contains several smooth effects. For such a complex model determining confidence intervals based on asymptotic assumptions might fail. Due to the vast increase of computational power Markov chain Monte Carlo (MCMC) simulations offer an attractive toolbox to provide valid credible intervals.

To be able to apply this technique to a GAM, the posterior distribution has to be formulated (Brezger and Lang 2006). MCMC samples of the posterior distribution can be efficiently generated by approximating a full-conditional distribution using a second order Taylor series expansion of the log-posterior centered at the last state (Gamerman 1997; Fahrmeir *et al.* 2013; Umlauf *et al.* 2017). Moreover, in most situations the structure of the sampling scheme reduces to an iteratively weighted least squares (IWLS) updating step for which highly efficient algorithms are available (Lang *et al.* 2014).

The ECMWF based models, selected by gradient boosting with stability selection, and the climatological baseline models are estimated by MCMC sampling. 1000 independent realizations of the regression coefficients are drawn from the Markov chains, which enables inference of the effects, predictions and out-of-sample scores.

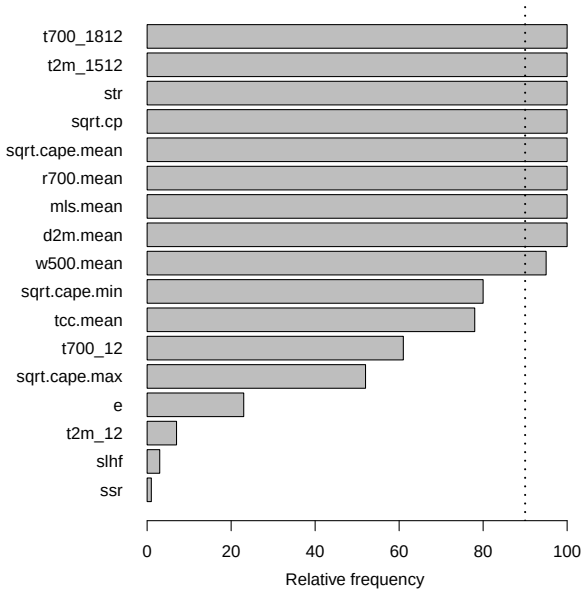

Figure 2: Results of the stability selection procedure for the model for day 1 and 64×64 km^2 resolution. The variable names on the y axis indicate that the associated nonlinear effect was selected. The vertical dotted line highlights the threshold of 90% for the final model.

4. Results

In this section the selection procedure is illustrated along one example case for one particular spatial resolution and lead time. Afterwards the predictive performance across the different resolutions and forecasts horizons is analyzed.

The selection procedure and the estimation of the final models is performed on data of four years (2010–2013), leaving data of two years (2014–2015) for evaluating the predictive performance of the final models.

Model selection

In total 18 models are fitted, i.e., five models with ECMWF covariates and one baseline model containing the climatological probability for each of the three spatial resolutions. The variable selection based on boosting and stability selection is performed for the 15 models with ECMWF covariates.

The results of the stability selection for the model referring to the resolution 64×64 km^2 and the forecast horizon of one day are visualized in Figure 2. The boosting algorithm was run on 100 distinct random subsamples of size $n/2$ from the training data until $q = 12$ terms were selected. The bars in Figure 2 indicate the relative frequency for a term $f_j(\mathbf{x}_j)$ being selected in the 100 boosting runs.

For example the term $f_{\texttt{t700\_1812}}$ was selected in each of the 100 runs. On the other side of the scale $f_{\texttt{ssr}}$ was selected only once. However, all 111 terms that are not listed on the y-axis have not been selected at all. Neither the seasonal term f_{time} nor the spatial term f_{space} were selected, which indicates that all variability over the considered time of the year and domain

can already be explained by the resulting effects from the ECMWF covariates.

All terms for which the relative frequency exceeds the threshold 90% (dotted line) enter the final model. Thus in this case the final model contains nine additive terms. For the given number of potential predictors $p = 128$ and the tuning parameters of the stability selection $q = 12$ and a threshold of 90%, this procedure ensures that the expected number of falsely included terms (non-informative effects) is less than 1 (cf. Eq. 6 in Hofner *et al.* 2015).

In order to provide inference for the effects of the final model a MCMC sampling is performed. Figure 3 shows the resulting effects for the model with the resolution 64×64 km^2 and the forecast horizon of one day. The effects are ordered according to their effect size, which is here defined as the absolute difference of the maximum and minimum value of the effect.

Mean relative humidity at 700 hPa (`r700.mean`) is the most influential covariate (Figure 3a). The absolute difference between the maximum and minimum value of the effect is 4.13 on the logit scale. Between the lower bound and 80% $f_{\texttt{r700.mean}}$ increases, which means that higher values of relative humidity in the ECMWF correspond to higher probabilities in the prediction of thunderstorms or lightning activity. However, at 80% the effect reaches a maximum and decreases slightly for higher values.

Other important effects are associated with the differences of temperature at 700 hPa between 1800 UTC and 1200 UTC (`t700_1812`), the square root of convective precipitation (`sqrt.cp`) and the proxy for mean layer stability (`mls.mean`, cf. Table 1). $f_{\texttt{t700\_1812}}$ and $f_{\texttt{mls.mean}}$ decrease both nonlinearly. The effect of the square root of the diagnostic variable convective precipitation is monotonic increasing.

The effect of surface net thermal radiation (Figure 3e) reveals a very interesting shape. It first increases from −0.17 to 0.89 at a value of $-1.78\ 10^6\ Jm^{-2}$, and decreases afterwards to −1.31 on the logit scale. However, on occasions with high absolute values of longwave heat fluxes (left hand side of the x-scale) the overall model would predict very small probabilities. This is due to a compensation effect between the additive terms. High absolute values of longwave heat fluxes coincide with low values of relative humidity at 700 hPa, for which $f_{\texttt{r700.mean}}$ is very small. In other words, if surface net thermal radiation would be employed as a single predictor, the increase on the left side of the scale would be more pronounced.

The effect of `d2m.mean` is monotonic increasing and spans a range of 2.08. The effects of `w500.min`, `sqrt.cape.mean` and `t2m_1512` are all less than unity on the logit scale.

This procedure—variable selection by combining gradient boosting and stability selection, and fitting the final model by MCMC sampling—was performed for all 15 models that build on ECMWF-HRES output. The effects presented for the example above are representative. All of these nine effects—except `d2m.mean` and `w500.mean`—were selected in a majority of models. In addition the effects of the mean of total cloud cover and the mean of CAPE were selected in more than 50% of the models.

The selection results can be summarized as follows. Increasing the resolution also increases the number of selected terms. For models with a longer forecast horizon, the number of selected terms decreases slightly. The median effect size decreases for increasing resolution as well as for increasing forecast horizons.

Predictive performance

The predictive performance was evaluated on the data from 2014 and 2015 by means of

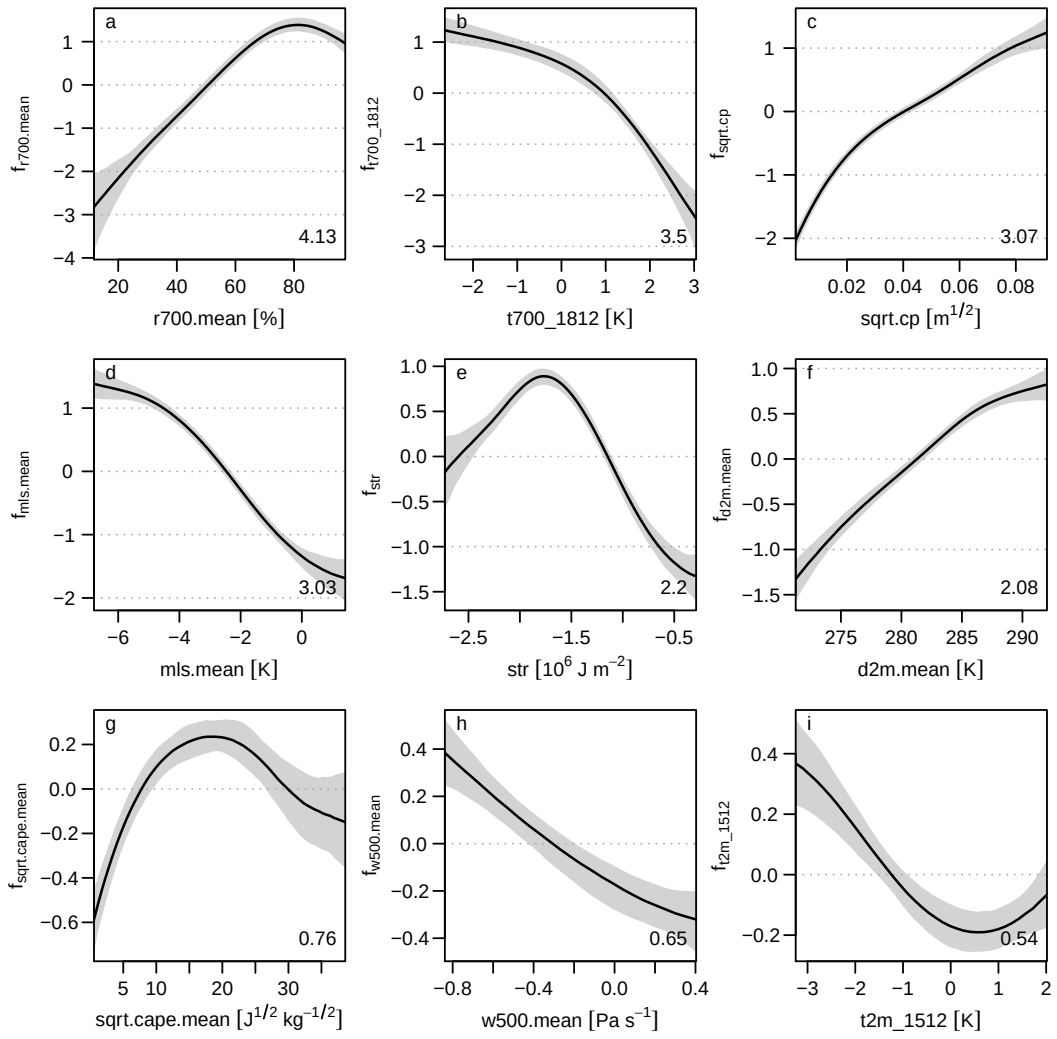

Figure 3: Effects and 95% credible intervals of the model for day 1 and 64×64 km^2 resolution estimated via MCMC sampling. The effects are displayed on the logit scale. The number in the bottom right corner of each panel indicates the absolute range of the effect. The x axes are cropped at the 1% and 99% quantile of the respective covariate to enhance graphical representation.

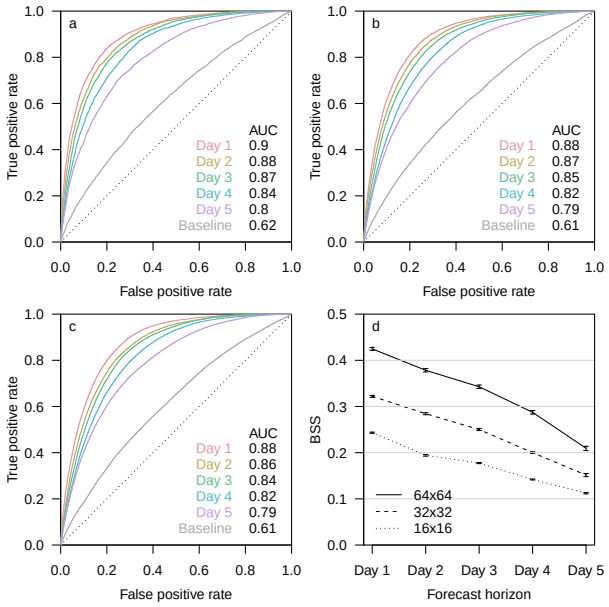

Figure 4: Predictive performance: Receiver operating characteristics diagrams for 64×64 km^2 (a), 32×32 km^2 (b) and 16×16 km^2 (c) resolution. AUC stands for the *area under curve*. (d) Brier skill score for all models with the baseline climatology as reference. The 95% intervals are derived from MCMC samples.

receiver operating characteristics (ROC) and Brier skill score (BSS). All scores show that the models based on ECMWF covariates are superior to the baseline models, i.e., climatologies.

Figure 4a–c shows the ROC diagram (Robin *et al.* 2011) for the models with the spatial resolutions 64×64 km^2 (a), 32×32 km^2 (b) and 16×16 km^2 (c). The diagram illustrates how well the predictions discriminate between lightning and no lightning. The curves for day 1 show that the probabilistic forecasts can be transformed to a binary prediction with a true positive rate greater than 80% and a false positive rate of less than 20%.

The area under curve (AUC) summarizes the receiver operating characteristics (Robin *et al.* 2011). The ECMWF based models are superior to the baseline models. The results for the different resolutions are comparable.

The BSS for all models is displayed in Figure 4d. The BSS is highest (0.42) for the coarsest resolution (64×64 km^2) and the shortest forecast horizon (1 day), and smallest (0.11) for the finest resolution (16×16 km^2) and longest forecast horizon (5 days). 95% intervals of the BSS were obtained using the MCMC samples.

All forecasts are well calibrated. The skill for all resolutions decreases from short to long forecast horizons, which is due to the decrease in sharpness of the forecasts, which will be discussed further in Section 5.

Finally, the spatial distribution of BSS for the model with the finest resolution and the longest forecast horizon, i.e., 16×16 km^2 and 5 days, is discussed (Figure 5), which is the model with the lowest overall skill (Figure 4d). MCMC samples were used to test at 5% level if BSS values are positive. Positive values mean that predictions from the post-processing are superior to the climatology. This is given around the Alps and in the northeastern part of the domain.

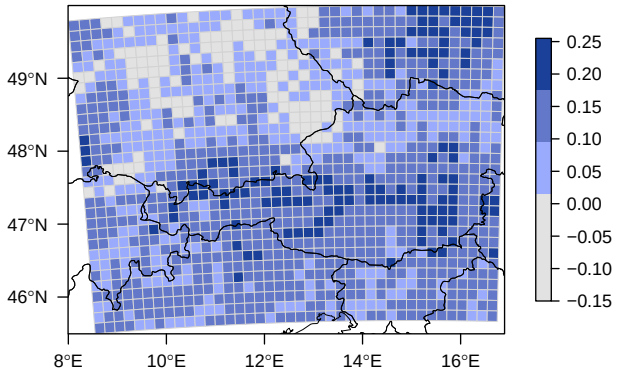

Figure 5: Spatial predictive performance: The Brier skill score for the model with 16×16 km^2 resolution and a forecast horizon of 5 days. All BSS greater than 0.015 (shown in blueish color) are significantly positive at the 5% level. Tests are based on MCMC samples.

If the BSS in a grid is not significantly positive the limit of predictability is reached. This is the case for regions further north of the Alps.

5. Discussion

One representative example (22 July 2015) is presented in order to highlight the information that can be gained from the introduced models.

The top left panel of Figure 6 shows the verifying observation for 22 July 2015, on the resolution 16×16 km^2, where ones (zeros) indicate cells in which lightning was (not) observed. The top mid panel shows the climatological probability for lightning activity in the cells for the same day compiled by the baseline model.

The baseline model reveals areas at the northern and southern rim of the Alps in which lightning activity is relatively likely with climatological probabilities ranging up to 26.5% on the southern rim. The lowest values around 10% can be found in the northern part of the domain. This pattern is in line with earlier studies (Feudale *et al.* 2013; Wapler 2013). The mean of the climatological probabilities for this day is 16.2%.

The bottom panels of Figure 6 illustrates how the predictions made by the GAMs with ECMWF predictors evolve from longer forecast horizons to shorter forecasts horizons. The bottom right panel shows the forecast with the model based on the ECMWF-HRES data with the lead times 108h, 111h and 114h, i.e., 5 days before 22 July. The mean of the predicted probabilities is 28.9%, and thus clearly above the climatological value. However, probabilities spread homogeneously over the domain with mid 50% of the values lying between 19.0% and 38.7%.

The spatial pattern of the forecast from 3 days before the event (bottom mid panel Figure 6) is already visible. There is a region with low values in the northwest of the domain, which can be distinguished from the rest of the domain with higher values.

The forecast made for the lead times 12h to 18h (bottom left panel) reveals sharp edges between the regions with high and low probabilities. The lower quarter of the predicted probabilities ranges from 0% to 1.6%, and the upper quarter from 59.3% to 83.7%. Thus the

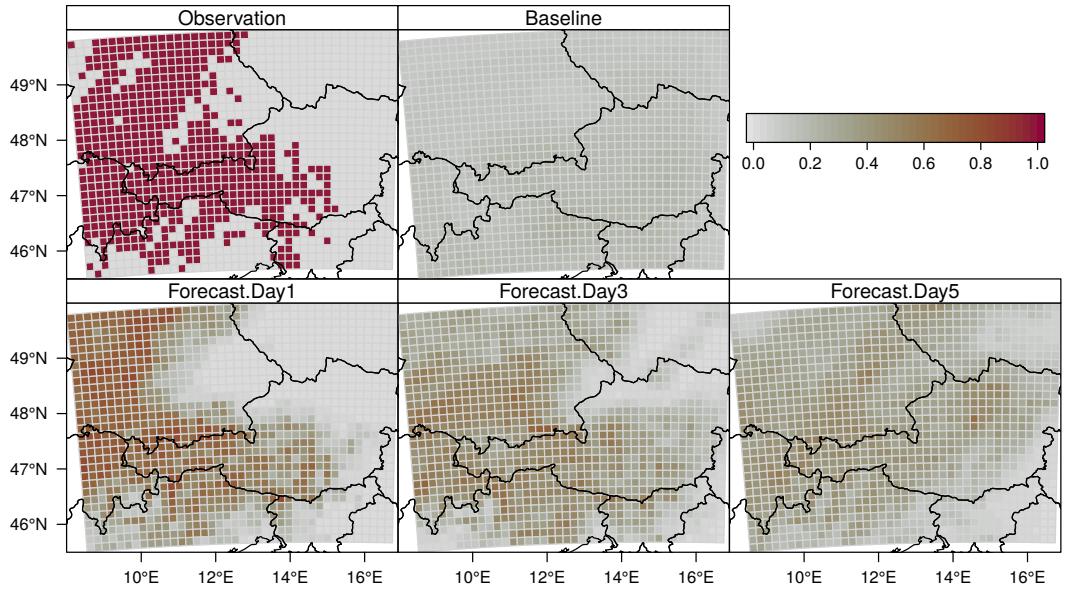

Figure 6: Spatial example: Top panel: Observed lightning activity for the afternoon (1200–1800 UTC) on 22 July 2015 and baseline climatology for the probability of thunderstorms for the same time interval. Bottom panel: Probabilistic forecasts for thunderstorms compiled 1, 3 and 5 days before 22 July 2015.

forecast provides a sharp information about the spatial pattern of the forthcoming weather event. In comparison with the verifying observation (top panel 6) the spatial pattern is well reproduced by the prediction.

For the same case (22 July 2015) the temporal evolution of predicted probabilities is highlighted for two sample locations, i.e., the grid cells associated with the airports of Zurich (ZRH) and Vienna (VIE). Figure 7 shows the probabilities for thunderstorms dependent on the forecast horizon.

Five days before the event probabilities of 36.1% and 32.8% were predicted for ZRH and VIE, respectively. These values are clearly greater than the corresponding climatological probabilities, 15.8% and 12.5%. When coming closer to the date of interest the probabilities for ZRH increase and for VIE decrease. For a forecast horizon of 3 days the predicted probability at VIE drops below the climatological one. For day 1 the predicted probability for VIE can not be distinguished from zero. The value for ZRH on the shortest forecast horizon is 75.6%. On 22 July 2015, lightning was observed in the grid cell containing ZRH, but not in VIE.

6. Conclusions

This study explores generalized additive models (GAM) and gradient boosting with stability selection as a tool for predicting thunderstorms by making use of numerical weather prediction (NWP) output. The Eastern Alps in Europe serve as study region. Observations of lightning strokes provide a proxy for the occurrence of thunderstorms. GAMs capture the potential nonlinear relationship between the covariates and the response while boosting with stability selection offers an objective way to select a stable subset of covariates and to control the

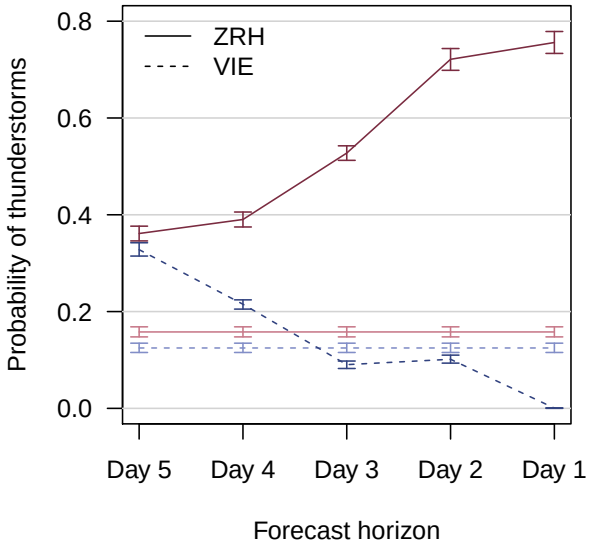

Figure 7: Temporal example for the grids in which the airports near Zurich (ZRH) and Vienna (VIE) are located (Figure 1). The predicted probabilities are connected with solid (ZRH) and dashed (VIE) lines, respectively. 95% intervals are based on MCMC samples. Climatologies are added in light colors.

number of falsely selected terms.

The resulting predictions are skillful to the longest evaluated forecast horizon of 5 days and the finest spatial resolution of 16×16 km^2.

Computational details

The statistical modeling has been carried out using the software environment R (R Core Team 2017). The add-on package **bamlss** (Umlauf *et al.* 2017) offers a flexible toolbox for complex regression models such as GAMs. It allows to perform gradient boosting via the model fitting engine function `boost()`, and to simulate MCMC samples of the posterior distribution with the engine function `GMCMC()`.

Acknowledgments

We acknowledge the funding of this work by the Austrian Research Promotion Agency (FFG) project *LightningPredict* (Grant No. 846620). The computational results presented have been achieved using the HPC infrastructure LEO of the University of Innsbruck.

References

Bertram I, Mayr GJ (2004). "Lightning in the Eastern Alps 1993–1999, Part I: Thunderstorm

Tracks." *Nat. Hazards Earth Syst. Sci.*, **4**(4), 501–511. doi:10.5194/nhess-4-501-2004.

Brezger A, Lang S (2006). "Generalized Structured Additive Regression Based on Bayesian P-Splines." *Comp. Stat. Data Anal.*, **50**(4), 967–991. doi:10.1016/j.csda.2004.10.011.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statist. Sci.*, **22**(4), 477–505. doi:10.1214/07-STS242.

Dabernig M, Mayr GJ, Messner JW, Zeileis A (2017). "Spatial Ensemble Post-Processing with Standardized Anomalies." *Quart. J. Roy. Meteor. Soc.*, **143**(703), 909–916. doi:10.1002/qj.2975.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-34333-9.

Feudale L, Manzato A, Micheletti S (2013). "A Cloud-to-Ground Lightning Climatology for North-Eastern Italy." *Adv. Sci. Res.*, **10**(1), 77–84. doi:10.5194/asr-10-77-2013.

Freund Y, Schapire RE (1995). *A Desicion-Theoretic Generalization of On-Line Learning and an Application to Boosting*, pp. 23–37. Springer, Berlin. doi:10.1007/3-540-59119-2_166.

Gamerman D (1997). "Sampling from the Posterior Distribution in Generalized Linear Mixed Models." *Stat. Comput.*, **7**(1), 57–68. ISSN 0960-3174. doi:10.1023/a:1018509429360.

Gijben M, Dyson LL, Loots MT (2017). "A Statistical Scheme to Forecast the Daily Lightning Threat over Southern Africa Using the Unified Model." *Atmos. Res.*, **194**, 78–88. doi:10.1016/j.atmosres.2017.04.022.

Hastie T, Tibshirani R (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton.

Hofner B, Boccuto L, Göker M (2015). "Controlling False Discoveries in High-Dimensional Situations: Boosting with Stability Selection." *BMC Bioinformatics*, **16**(1). doi:10.1186/s12859-015-0575-3.

Houze RA (2014). *Cloud Dynamics*, volume 104 of *International Geophysics*. Academic Press. doi:10.1016/B978-0-12-374266-7.00003-2.

Lang S, Umlauf N, Wechselberger P, Harttgen K, Kneib T (2014). "Multilevel Structured Additive Regression." *Stat. Comput.*, **24**(2), 223–238. ISSN 0960-3174. doi:10.1007/s11222-012-9366-0.

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012). "Generalized Additive Models for Location, Scale and Shape for High Dimensional Data—A Flexible Approach based on Boosting." *J. Roy. Stat. Soc. C*, **61**(3), 403–427. doi:10.1111/j.1467-9876.2011.01033.x.

Meinshausen N, Bühlmann P (2010). "Stability Selection." *J. Roy. Stat. Soc. B*, **72**(4), 417–473. doi:10.1111/j.1467-9868.2010.00740.x.

Messner JW, Mayr GJ, Zeileis A (2017). "Nonhomogeneous Boosting for Predictor Selection in Ensemble Post-Processing." *Mon. Wea. Rev.*, **145**(1), 137–147. doi:10.1175/MWR-D-16-0088.1.

Miller A (2002). *Subset Selection in Regression*, volume 95 of *Monographs on Statistics and Applied Probability*. 2nd edition. Chapman & Hall/CRC, Boca Raton.

Piper D, Kunz M (2017). "Spatiotemporal Variability of Lightning Activity in Europe and the Relation to the North Atlantic Oscillation Teleconnection Pattern." *Nat. Hazards Earth Syst. Sci.*, **17**(8), 1319–1336. doi:10.5194/nhess-17-1319-2017.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011). "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves." *BMC Bioinformatics*, **12**(1), 77. doi:10.1186/1471-2105-12-77.

Schmeits MJ, Kok KJ, Vogelezang DHP, van Westrhenen RM (2008). "Probabilistic Forecasts of (Severe) Thunderstorms for the Purpose of Issuing a Weather Alarm in the Netherlands." *Wea. Forecasting*, **23**(6), 1253–1267. doi:10.1175/2008WAF2007102.1.

Schulz W, Cummins K, Diendorfer G, Dorninger M (2005). "Cloud-to-Ground Lightning in Austria: A 10-Year Study Using Data from a Lightning Location System." *J. Geophys. Res.*, **110**(D9). doi:10.1029/2004JD005332.

Simon T, Umlauf N, Zeileis A, Mayr GJ, Schulz W, Diendorfer G (2017). "Spatio-Temporal Modelling of Lightning Climatologies for Complex Terrain." *Nat. Hazards Earth Syst. Sci.*, **17**(3), 305–314. doi:10.5194/nhess-17-305-2017.

Stauffer R, Umlauf N, Messner JW, Mayr GJ, Zeileis A (2017). "Ensemble Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies." *Mon. Wea. Rev.*, **145**(3), 955–969. doi:10.1175/MWR-D-16-0260.1.

Umlauf N, Klein N, Zeileis A (2017). "BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond)." *J. Comput. Graph. Stat.*, **0**(ja), 0–0. doi:10.1080/10618600.2017.1407325.

Wahl S (2015). *Uncertainty in Mesoscale Numerical Weather Prediction: Probabilistic Forecasting of Precipitation*. Rheinische Friedrich-Wilhelms-Universität Bonn. Available online at http://hss.ulb.uni-bonn.de/2015/4190/4190.htm.

Wapler K (2013). "High-Resolution Climatology of Lightning Characteristics within Central Europe." *Meteor. Atmos. Phys.*, **122**(3-4), 175–184. doi:10.1007/s00703-013-0285-1.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Affiliation:

Thorsten Simon
Institute of Atmospheric and Cryospheric Sciences
University of Innsbruck
Innrain 52f
6020 Innsbruck, Austria
E-mail: Thorsten.Simon@uibk.ac.at

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://uibk.ac.at/eeecon/wopec/

2017-25 **Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Probabilistic Forecasting of Thunderstorms in the Eastern Alps

2017-24 **Florian Lindner:** Choking under pressure of top performers: Evidence from Biathlon competitions

2017-23 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood

2017-22 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Forecasting low-visibility procedure states with tree-based statistical methods

2017-21 **Philipp Kneringer, Sebastian J. Dietz, Georg J. Mayr, Achim Zeileis:** Probabilistic nowcasting of low-visibility procedure states at Vienna International Airport during cold season

2017-20 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior

2017-19 **Martin Geiger, Richard Hule:** The role of correlation in two-asset games: Some experimental evidence

2017-18 **Rudolf Kerschbamer, Daniel Neururer, Alexander Gruber:** Do the altruists lie less?

2017-17 **Meike Köhler, Nikolaus Umlauf, Sonja Greven:** Nonlinear association structures in flexible Bayesian additive joint models

2017-16 **Rudolf Kerschbamer, Daniel Muller:** Social preferences and political attitudes: An online experiment on a large heterogeneous sample

2017-15 **Kenneth Harttgen, Stefan Lang, Judith Santer, Johannes Seiler:** Modeling under-5 mortality through multilevel structured additive regression with varying coefficients for Asia and Sub-Saharan Africa

2017-14 **Christoph Eder, Martin Halla:** Economic origins of cultural norms: The case of animal husbandry and bastardy

2017-13 **Thomas Kneib, Nikolaus Umlauf:** A Primer on Bayesian Distributional Regression

2017-12 **Susanne Berger, Nathaniel Graham, Achim Zeileis:** Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R

2017-11 **Natalia Danzer, Martin Halla, Nicole Schneeweis, Martina Zweimüller:** Parental leave, (in)formal childcare and long-term child outcomes

2017-10 **Daniel Muller, Sander Renes:** Fairness views and political preferences - Evidence from a large online experiment

2017-09 **Andreas Exenberger:** The Logic of Inequality Extraction: An Application to Gini and Top Incomes Data

2017-08 **Sibylle Puntscher, Duc Tran Huy, Janette Walde, Ulrike Tappeiner, Gottfried Tappeiner:** The acceptance of a protected area and the benefits of sustainable tourism: In search of the weak link in their relationship

2017-07 **Helena Fornwagner:** Incentives to lose revisited: The NHL and its tournament incentives

2017-06 **Loukas Balafoutas, Simon Czermak, Marc Eulerich, Helena Fornwagner:** Incentives for dishonesty: An experimental study with internal auditors

2017-05 **Nikolaus Umlauf, Nadja Klein, Achim Zeileis:** BAMLSS: Bayesian additive models for location, scale and shape (and beyond)

2017-04 **Martin Halla, Susanne Pech, Martina Zweimüller:** The effect of statutory sick-pay on workers' labor supply and subsequent health

2017-03 **Franz Buscha, Daniel Müller, Lionel Page:** Can a common currency foster a shared social identity across different nations? The case of the Euro.

2017-02 **Daniel Müller:** The anatomy of distributional preferences with group identity

2017-01 **Wolfgang Frimmel, Martin Halla, Jörg Paetzold:** The intergenerational causal effect of tax evasion: Evidence from the commuter tax allowance in Austria

2016-33 **Alexander Razen, Stefan Lang, Judith Santer:** Estimation of spatially correlated random scaling factors based on Markov random field priors

2016-32 **Meike Köhler, Nikolaus Umlauf, Andreas Beyerlein, Christiane Winkler, Anette-Gabriele Ziegler, Sonja Greven:** Flexible Bayesian additive joint models with an application to type 1 diabetes research

2016-31 **Markus Dabernig, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Simultaneous ensemble post-processing for multiple lead times with standardized anomalies

2016-30 **Alexander Razen, Stefan Lang:** Random scaling factors in Bayesian distributional regression models with an application to real estate data

2016-29 **Glenn Dutcher, Daniela Glätzle-Rützler, Dmitry Ryvkin:** Don't hate the player, hate the game: Uncovering the foundations of cheating in contests

2016-28 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Tricks for improving non-homogeneous regression for probabilistic precipitation forecasts: Perfect predictions, heavy tails, and link functions

2016-27 **Michael Razen, Matthias Stefan:** Greed: Taking a deadly sin to the lab

2016-26 **Florian Wickelmaier, Achim Zeileis:** Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models

2016-25 **Michel Philipp, Carolin Strobl, Jimmy de la Torre, Achim Zeileis:** On the estimation of standard errors in cognitive diagnosis models

2016-24 **Florian Lindner, Julia Rose:** No need for more time: Intertemporal allocation decisions under time pressure

2016-23 **Christoph Eder, Martin Halla:** The long-lasting shadow of the allied occupation of Austria on its spatial equilibrium

2016-22 **Christoph Eder:** Missing men: World War II casualties and structural change

2016-21 **Reto Stauffer, Jakob Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies *published in Monthly Weather Review*

2016-20 **Christina Bannier, Eberhard Feess, Natalie Packham, Markus Walzl:** Incentive schemes, private information and the double-edged role of competition for agents

2016-19 **Martin Geiger, Richard Hule:** Correlation and coordination risk

2016-18 **Yola Engler, Rudolf Kerschbamer, Lionel Page:** Why did he do that? Using counterfactuals to study the effect of intentions in extensive form games

2016-17 **Yola Engler, Rudolf Kerschbamer, Lionel Page:** Guilt-averse or reciprocal? Looking at behavioural motivations in the trust game

2016-16 **Esther Blanco, Tobias Haller, James M. Walker:** Provision of public goods: Unconditional and conditional donations from outsiders

2016-15 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** Predictive bookmaker consensus model for the UEFA Euro 2016

2016-14 **Martin Halla, Harald Mayr, Gerald J. Pruckner, Pilar García-Gómez:** Cutting fertility? The effect of Cesarean deliveries on subsequent fertility and maternal labor supply

2016-13 **Wolfgang Frimmel, Martin Halla, Rudolf Winter-Ebmer:** How does parental divorce affect children's long-term outcomes?

2016-12 **Michael Kirchler, Stefan Palan:** Immaterial and monetary gifts in economic transactions. Evidence from the field

2016-11 **Michel Philipp, Achim Zeileis, Carolin Strobl:** A toolkit for stability assessment of tree-based learners

2016-10 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** Affirmative action or just discrimination? A study on the endogenous emergence of quotas *published in Journal of Economic Behavior and Organization*

2016-09 **Loukas Balafoutas, Helena Fornwagner:** The limits of guilt

2016-08 **Markus Dabernig, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Spatial ensemble post-processing with standardized anomalies

2016-07 **Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

2016-06 **Michael Razen, Jürgen Huber, Michael Kirchler:** Cash inflow and trading horizon in asset markets

2016-05 **Ting Wang, Carolin Strobl, Achim Zeileis, Edgar C. Merkle:** Score-based tests of differential item functioning in the two-parameter model

2016-04 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Non-homogeneous boosting for predictor selection in ensemble post-processing

2016-03 **Dietmar Fehr, Matthias Sutter:** Gossip and the efficiency of interactions

2016-02 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Rankings and risk-taking in the finance industry

2016-01 **Sibylle Puntscher, Janette Walde, Gottfried Tappeiner:** Do methodical traps lead to wrong development strategies for welfare? A multilevel approach considering heterogeneity across industrialized and developing countries

University of Innsbruck

Working Papers in Economics and Statistics

2017-25

Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis
Probabilistic Forecasting of Thunderstorms in the Eastern Alps

Abstract
A probabilistic forecasting method to predict thunderstorms in the European Eastern Alps is developed. A statistical model links lightning occurrence from the ground-based ALDIS detection network to a large set of direct and derived variables from a numerical weather prediction (NWP) system. The NWP system is the high resolution run (HRES) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The statistical model is a generalized additive model (GAM) framework, which is estimated by Markov chain Monte Carlo (MCMC) simulation. Gradient boosting with stability selection serves as a tool for selecting a stable set of potentially nonlinear terms. Three grids from 64Ã—64 km 2 to 16Ã—16 km 2 and 5 forecasts horizons from 5 to 1 day ahead are investigated to predict thunderstorms during afternoons (1200 UTC to 1800 UTC). Frequently selected covariates for the nonlinear terms are variants of convective precipitation, convective potential available energy, relative humidity and temperature in the mid layers of the troposphere, among others. All models, even for a lead time of five days, outperform a forecast based on climatology in an out-of-sample comparison. An example case illustrates that coarse spatial patterns are already successfully forecast five days ahead.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)